ANIMAL COMMUNICATION

Extensive compositionality in the vocal system of bonobos

M. Berthet[1,2,3]*, M. Surbeck[4,5]†, S. W. Townsend[1,2,3,6]†

Compositionality, the capacity to combine meaningful elements into larger meaningful structures, is a hallmark of human language. Compositionality can be trivial (the combination's meaning is the sum of the meaning of its parts) or nontrivial (one element modifies the meaning of the other element). Recent studies have suggested that animals lack nontrivial compositionality, representing a key discontinuity with language. In this work, using methods borrowed from distributional semantics, we investigated compositionality in wild bonobos and found that not only does each call type of their repertoire occur in at least one compositional combination, but three of these compositional combinations also exhibit nontrivial compositionality. These findings suggest that compositionality is a prominent feature of the bonobo vocal system, revealing stronger parallels with human language than previously thought.

A quintessential feature of human language is the capacity to combine elements. For example, morphemes can be combined into words (e.g., "bio" + "logy" = "biology") or words into sentences ("Biology is interesting"). This is possible because of compositionality, whereby meaningful units are combined into larger structures whose meaning is determined by the meanings of the parts and the way they are combined (*1*, *2*).

Compositionality can take two forms. In its "trivial" (or "intersective") version, each element of the combination contributes to the meaning of the whole independently of the other element, and the combination is interpreted by the conjunction of its parts (*3*–*5*). For example, "blond dancer" refers to a person who is both blond and a dancer; if this person is also a doctor, we can infer that they are a blond doctor as well. However, compositional syntax can also be "nontrivial" (or "nonintersective"): The units constituting a combination do not contribute independent meaning, but instead, they combine so that one part of the combination modifies the other (*3*, *4*). For example, the meaning of the expression "bad dancer" does not refer to a bad person who is also a dancer. Indeed, if this person is also a doctor, we cannot infer that they are a bad doctor. Here, "bad" does not have a meaning independent from "dancer"; rather, it complements it (*3*–*5*).

Compositionality as a phenomenon might not be unique to human language. Numerous studies in birds and primates have demonstrated that animals are capable of combining meaningful vocalizations into trivially compositional structures (*6*–*8*). To our knowledge, however, unambiguous evidence of nontrivial compositionality in animals from systematically collected quantitative data is still lacking (*3*, *9*–*16*).

In this work, we provide robust empirical evidence for the presence of nontrivial compositionality in wild bonobos (*Pan paniscus*). First, we leveraged a framework established by Berthet *et al.* (*2*) that investigates meaning by considering all aspects of context that co-occur with the emission of the signal. This approach defines the meaning of a signal as the set of features of circumstances (FoCs) that appear at a rate greater than chance across the signal's occurrences (*2*). We recorded 700 wild bonobo calls and call combinations (hereafter, vocal utterances; table S1) and systematically collected more than 300 FoCs for each utterance (see materials and methods and table S2). Second, using a method adapted from distributional semantics, a linguistic approach that quantifies meaning similarities between words (*17*), we used these FoCs to map bonobo utterance types within a multidimensional space (hereafter, semantic space) and quantify meaning similarities between utterance types.

Last, to investigate whether the bonobo call system is compositional, we applied a multistep process previously used by Trujillo and Holler to identify nontrivial compositionality in human multimodal communication (*12*). Under this approach, a combination AB is considered compositional if (i) the meaning of A is distinct from that of B, (ii) the meaning of AB is different from that of A and that of B, and (iii) the meaning of AB is derived from the meaning of A and B. The fourth step determines whether a compositional combination is trivial or nontrivial. More precisely, the combination AB represents a nontrivial compositional structure if it is compositional [i.e., it fulfils criteria (i) to (iii)] and if (iv) the meaning of AB is different from the meaning of A+B (see fig. S1 for an explicit visualization of the process). Our study shows that all seven bonobo call types that were considered in our analysis combine into four compositional structures, of which three exhibit nontrivial compositionality.

[1]Department of Evolutionary Anthropology, University of Zürich, Zürich, Switzerland. [2]Department of Comparative Language Sciences, University of Zürich, Zürich, Switzerland. [3]Institute for the Interdisciplinary Study of Language Evolution, University of Zürich, Zürich, Switzerland. [4]Department of Human Evolutionary Biology, Harvard University, Cambridge, MA, USA. [5]Department of Human Behavior, Ecology and Culture, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany. [6]Department of Psychology, University of Warwick, Coventry, UK.
*Corresponding author. Email: melissa.berthet.ac@gmail.com
†These authors contributed equally to this work.

Results

Semantic space of bonobo utterances

We adapted a distributional semantics framework to investigate the meaning of bonobos' call combinations in relation to single calls. Distributional semantics is based on the distributional hypothesis, which states that words with close meanings are used in similar contexts (*17*). Distributional semantics represents word meaning in a semantic space by converting each word into a vector so that relationships between words can be quantified using geometric relations (*18*). We estimated similarities between single call types and call combinations of wild bonobos using a multiple correspondence analysis (MCA) performed on the FoCs. The MCA is similar to a principal components analysis (PCA) but is conducted on categorical data: It performs a dimension reduction and then quantifies the statistical relationship between a specific utterance type and several FoCs (*19*) (see materials and methods). We found that the first five dimensions of the MCA explained about 24% of the variance in the data (15.00 + 3.12 + 2.46 + 2.18 + 2.14%). These five dimensions were used to create a semantic space in which each utterance was mapped using five-dimensional coordinates (Fig. 1 and figs. S2 to S4). Utterances with a similar meaning are closer to each other in the semantic space than utterances with a different meaning. These five-dimensional coordinates were used for the remaining analyses.

Compositionality analysis

After mapping the meaning of each call type and each combination, we investigated whether the combinations were nontrivially compositional. For this, we used a step-by-step process described in Trujillo and Holler (*12*).

Chance value

For each call type, we calculated a chance value representing the minimum Euclidian distance required to conclude that another utterance type has a different meaning from that call type. This later allowed us to determine whether distances between call types and/or combinations are greater than expected compared with the natural variance of the meaning of the single calls (table S3 and materials and methods).

(i) The meaning of A is different from the meaning of B

We investigated whether the single calls had significantly different meanings from each other. To do so, we calculated, for each possible pair of single calls, the difference between (i) the Euclidian distance separating the single calls

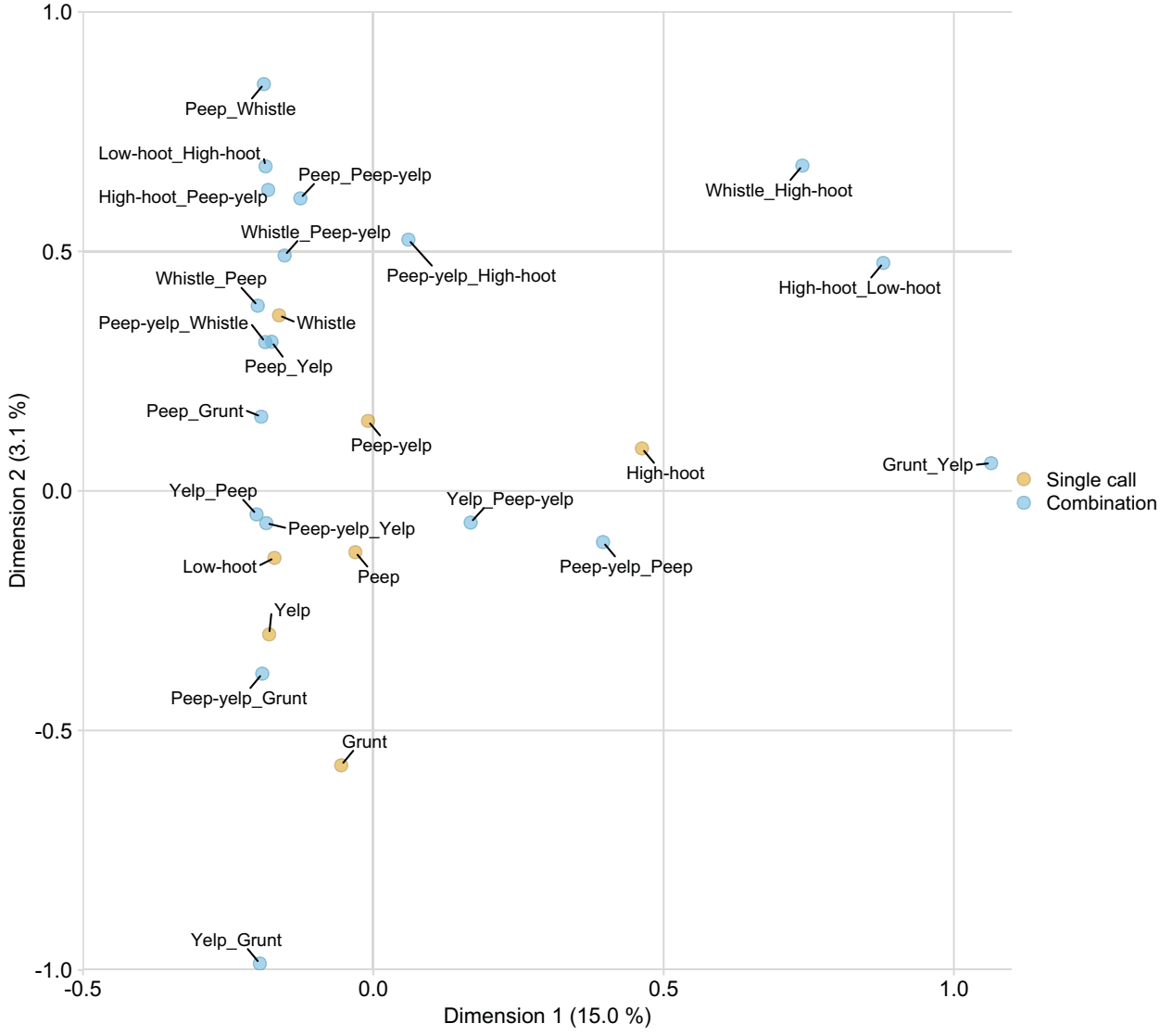

Fig. 1. First two dimensions of the semantic space of bonobos' vocal repertoire. Call types and combinations are plotted in a semantic space based on their inferred meaning (i.e., the FoCs associated with their emission). Utterances that have a similar meaning (i.e., that are emitted in similar contexts) are closer to each other than utterances emitted in different contexts. Combinations (in blue) are in the form A_B.

and (ii) the mean of the chance value of the calls. We then assessed whether this difference was overall significantly different from zero using a one-sample t test.

The meanings of call types were significantly different from each other ($t_{20} = 4.63$, $p < 0.001$). Specifically, all call types had a different meaning from each other except for three: Low-hoot had a similar meaning to Peep-yelp and Yelp (Fig. 2) because the difference between the pairs of call types and the mean chance value of the calls was less than zero (table S4).

(ii) The meaning of AB is different from the meaning of A and the meaning of B

We then investigated whether the call combinations had a different meaning from that of their constituents. To this end, we calculated, for each possible pair of one combination and one constituting call type of that combination, the difference between (i) the Euclidian distance separating the combination and its constituting call type, and (ii) the chance value of the constituting call type. We then assessed whether this difference was overall significantly different from zero using a one-sample t test.

The meaning of combinations was significantly different from the meaning of their constituents ($t_{37} = 7.36$, $p < 0.001$). Specifically, all combinations had a meaning different from their constituting call types, except for two: Peep-yelp_Peep and Peep-yelp_Yelp (Fig. 3 and table S5).

(iii) The meaning of AB is derived from the meaning of A and B

We investigated whether the meaning of the combinations was derived from the meaning of their elements. Following Trujillo and Holler (*12*), we calculated whether the meaning of a combination was closer to that of its constituent elements than any other element. Stated another way, we calculated whether the Euclidean distance between a combination and the call types constituting it was smaller than the distance between a combination and any other call type. For this, we calculated the pairwise Euclidian distance separating every single call from every combination. We then ran a linear model with distance as the dependent variable, and an interaction between the combination type (e.g., Yelp_Grunt) and whether the single call was a part of the combination (here termed the "relationship variable") as independent variables.

The meaning of combinations was closer to that of their constituents at the repertoire level: A likelihood ratio test detected a difference between the full model and its reduced version

(i.e., the model without the relationship variable) ($F_{19,95}$ = 2.13, p < 0.01). Specifically, posthoc analyses revealed that, for four combinations, the distance separating the combination from its constituting call types was smaller than the distance separating the combinations from call types that do not constitute it: High-hoot_Low-hoot (t_{95} = 2.07, p = 0.04), Peep_Whistle (t_{95} = 2.70, p < 0.01), Peep-yelp_High-hoot (t_{95} = 2.17, p = 0.03), and Yelp_Grunt (t_{95} = 3.17, p < 0.01) (Fig. 4 and table S6). Because these four combinations fulfil criteria (i) to (iii), they are all considered compositional.

(iv) The meaning of AB is different from the meaning of A+B

Finally, we investigated whether the four aforementioned compositional combinations were explained by trivial or nontrivial compositionality (*12*). We expected combinations exhibiting nontrivial compositionality to have a meaning different from adding the meaning of their parts. Following Trujillo *et al.* (*12*), we built a hypothetical additive combination by adding the coordinates of the constituting single call types. We then calculated the difference between (i) the Euclidian distance separating the combination and the additive combination and (ii) the mean chance value of the constituting call types of that combination. If that difference was greater than zero, we concluded that the combination exhibited nontrivial compositionality. Stated another way, we added the coordinates of A and B to create an A+B vector and calculated whether the Euclidean distance between A+B and AB was different from the mean chance value of A and B.

For all compositional combinations except Yelp_Grunt, the difference was greater than zero (High-hoot_Low-hoot: 0.42; Peep_Whistle: 0.36; Peep-yelp_High-hoot: 0.21). For Yelp_Grunt, the distance between the combination and an additive combination was not greater than the mean chance value (−0.02). That is, Yelp_Grunt exhibited trivial compositionality, whereas High-hoot_Low-hoot, Peep_Whistle, and Peep-yelp_High-hoot exhibited nontrivial compositionality.

Meaning of the utterances

In the final step, we derived a tentative meaning for the single calls and the combinations to illustrate how single calls combine into structures whose meaning is derived from the meaning of their parts. We extracted the meaning of each utterance type by determining which set of FoCs is associated with the utterance type at the repertoire level (see materials and

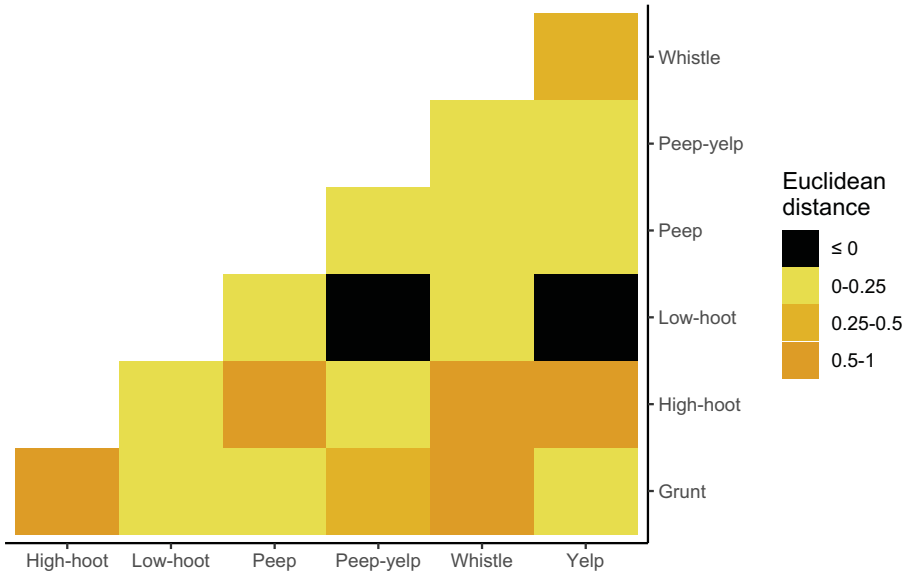

Fig. 2. Meaning differences between call types. The meaning difference is represented by whether the Euclidean distance between two call types is different from the mean of the chance value of these call types. Call types that overlap in meaning have a Euclidian distance of zero or less (in black).

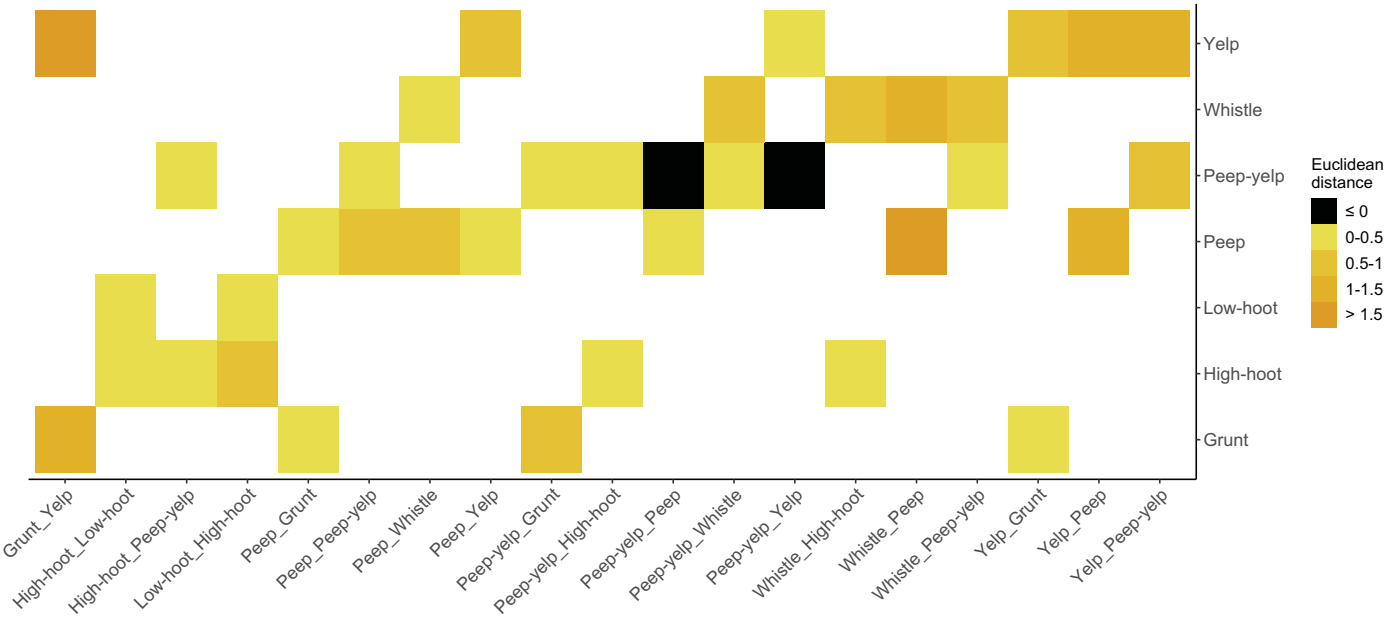

Fig. 3. Meaning differences between a combination A_B and its constituting call types A and B. Meaning differences between a combination A_B (e.g., Grunt_Yelp) (*x* axis) and its constituting call types A (e.g., Grunt) and B (e.g., Yelp) (*y* axis), represented as whether the Euclidean distance between the combination and a constituting call type is different from the chance value of this call type. Combinations whose meaning overlaps with one of its constituting call types have a Euclidian distance of zero or less with that call type (in black).

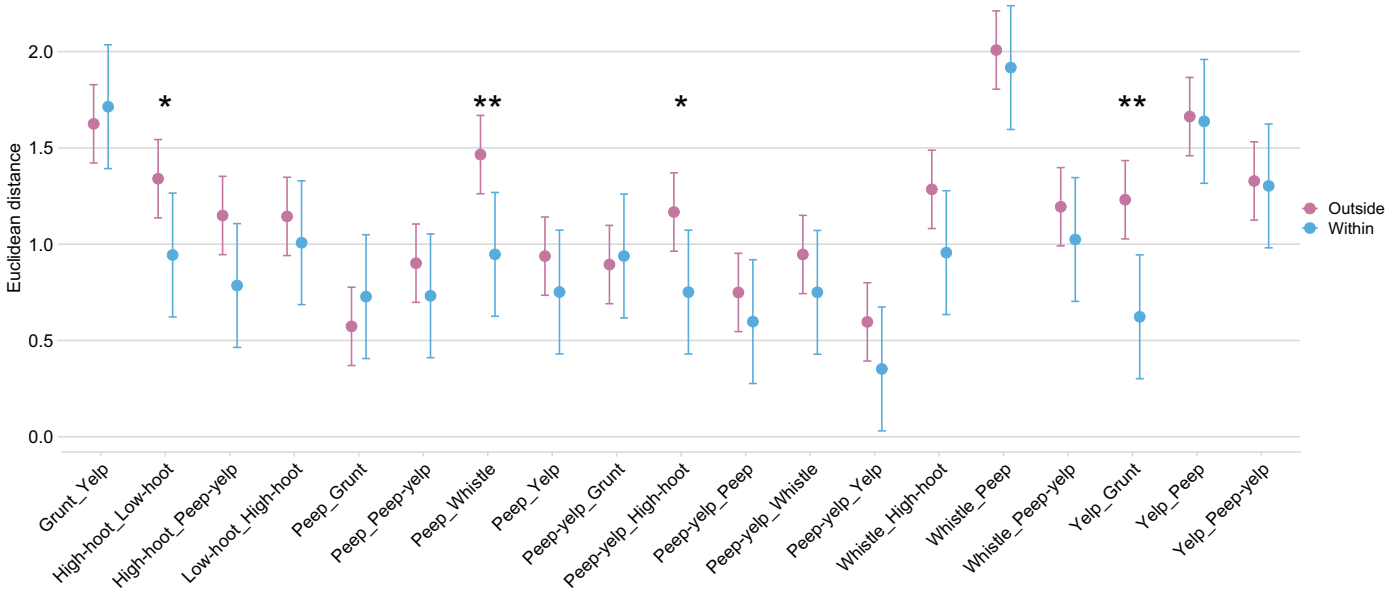

Fig. 4. Meaning difference between a combination A_B and single call types of the repertoire. Distance between the meaning of a given combination and that of its constituting call types (in blue) and between the meaning of a combination and that of call types that do not constitute it (in pink). Boxplots indicate the lower and higher confidence intervals, and dots indicate the estimated marginal means of the model. The symbols indicate significant differences (*$0.01 < p < 0.05$; **$p < 0.01$).

methods). Results are summarized in tables S7 and S8.

The semantic analysis of the Grunt suggests that it signals the ongoing activity of the caller. Because Grunts are given in a large variety of contexts (during grooming, resting, moving, feeding), they may signal that others should look at the caller to coordinate activities ("Look at me," in close interactions). High-hoots are emitted in different contexts and seem to signal the presence or location of the caller, especially in long-distance settings or dangerous situations ("Pay attention to me," for a larger audience or more considerable distances than the Grunts). The Low-hoots seem related to high-arousal situations, such as the construction of the night nest (*20*) ("I am excited"). The semantic analysis suggests that Yelps and Peeps have similar meanings, used as imperatives to coordinate social activities. However, Peeps seem to be more used as suggestions rather than strong imperatives: They are less associated with changes in behaviors in the audience and sometimes elicit vocal responses, which could suggest a quorum-based or "voting-like" system ("I would like to…"). Conversely, Yelps are associated with changes in the behaviors of the audience and no vocal response, suggesting that Yelps are more of a rigid imperative ("Let's do that"). The meaning of the Peep-yelp also varies with the context. It is mainly given during group encounters or when building night nests, maybe to coordinate intergroup interactions and encourage other parties to join ("Join!"). Finally, the Whistles seem to coordinate the spatial cohesion of the group ("Let's stay together").

The Yelps ("Let's do that") and Grunts ("Look at me") can be combined into the trivial compositional structure Yelp_Grunt, which seems to incite others to build a night nest. The High-hoot ("Pay attention to me") can be combined with the Low-hoot ("I am excited") into a nontrivial compositional combination High-hoot_Low-hoot, which is used when another individual is displaying, maybe to elicit a reaction in the audience (recruitment, "Pay attention to me because I am in distress") or to stop the display behavior of the other individual. The Peep ("I would like to…") can be combined with the Whistle ("Let's stay together") into a nontrivial compositional combination, Peep_Whistle, which is used in sensitive social contexts, for example, during copulations or displays, maybe to attract attention or assert rank. Finally, the Peep-yelp ("Join!") can be combined with the High-hoot ("Pay attention to me") into the nontrivial compositional structure Peep-yelp_High-hoot, which seems to be used to coordinate with other parties before traveling.

Discussion

In this study, we assessed whether bonobos combine calls into larger compositional structures. Using a holistic investigation of meaning derived from distributional semantics combined with a method for identifying compositionality in multimodal human language (*12*), we found that bonobos can combine all their calls into four compositional structures, of which one is a case of trivial compositionality (Yelp_Grunt) and three (High-hoot_Low-hoot, Peep_Whistle, Peep-yelp_High-hoot) are cases of nontrivial compositionality.

Our findings have three important implications. In humans, compositionality is essential for generativity, a crucial hallmark of language by which speakers can combine a finite set of elements into an infinite number of combinations that others can understand (*21*). Although most investigations of the compositional capacities of nonhuman species have been limited to one specific combination of a vocal system [for example, (*6*, *7*, *22*, *23*)], our study suggests that, in bonobos, this capacity is not restricted to a few isolated combinations. Indeed, each of the seven call types we investigated here represents a building block of a compositional structure. The first implication of our work is therefore that, akin to human language, compositionality is a pervasive component of bonobos' vocal communication.

Second, our data provide robust empirical evidence that nonhuman animals engage in nontrivial compositionality. In human language, nontrivial compositionality is particularly important in facilitating generativity by allowing meaningful elements to be combined into novel, nonadditive structures like "bad dancer," where the meaning of the whole is more than the meaning of its parts (*12*). Although a number of studies have demonstrated that animals can also combine meaningful vocalizations compositionally [see review in (*8*)], to our knowledge, no study to date has empirically shown that animals engage in complex nontrivial compositionality akin to that observed in human language (*11*, *13*, *24*). Our results indicate that nontrivial compositionality is not limited to humans and that bonobos, our closest living relative, also engage in nontrivial compositionality.

The extent to which our findings can inform the evolutionary roots of linguistic compositionality is yet to be determined. One interpretation of

the data could be that nontrivial compositionality can be traced as far back as the last common ancestor of bonobos and humans, 7 million to 13 million years ago (*25*). It is also possible that the absence of evidence in other species has been hampered by the lack of appropriate methods. Indeed, as we show here, detecting (nontrivial) compositionality largely depends on the capacity to reliably assess meaning and detect compositionality across an entire repertoire. Hence, a third important implication of our work is that we present a method for reliably inferring the meaning of all the signals of an animal's repertoire with minimal human judgment. This method can theoretically be applied to any taxa and any communication system (e.g., gestures, facial expressions, multimodal signals) to build a systematic survey of meaning and compositionality across species, ultimately allowing for a comprehensive investigation of the evolutionary pressures driving the evolution of these capacities.

It is worth noting that our study does not come without limitations, which might explain why the meaning of some utterance types (e.g., Low-hoots or Yelp_Grunts) are difficult to translate into human-relevant concepts. First, some meanings might be emotionally expressive (i.e., conveying information about the caller's emotional states), and future research should include FoCs related to the caller's emotional state, using direct emotion measurements such as infrared thermography (*26*). Second, our vocal repertoire might not be of sufficient granularity to capture the full complexity of the system: If some call types have subtle acoustic variants that convey different meanings, we may have misclassified them as one single call type, preventing reliable determination of their meaning. Finally, similar to birdsongs, some call types might have no meaning (*2*), which is not accounted for in our approach; for instance, the High-hoot might function to localize the caller rather than convey a specific message.

Although our focus is on vocal combinations, it is important to note that great apes also combine signals from different modalities, exposing the receiver to an array of additional multimodal information that may refine or modify a signal's meaning (*27*–*31*). For example, bonobos combine Contest-hoots with different gestures that potentially disambiguate the meaning of these vocalizations, that is, whether they intend to play with or challenge another individual (*32*). The extent to which combining vocalizations with other signals also modifies meaning in potentially nontrivial ways remains unknown, but it represents a critical follow-up research avenue to this study (*33*, *34*).

Our results (in particular, Fig. 1 and figs. S2 to S4) suggest that bonobos use call combinations in semantic areas in which single calls do not occur. That is, bonobos seem to combine calls to convey meaning that cannot be conveyed through single calls alone. Testing this hypothesis in the future will help us understand the selective forces that have given rise to compositionality in both animal communication and human language.

REFERENCES AND NOTES

1. B. H. Partee, in *Compositionality in Formal Semantics: Selected Papers* (Blackwell Publishing, 2004), pp. 153–181.
2. M. Berthet, C. Coye, G. Dezecache, J. Kuhn, *Biol. Rev. Camb. Philos. Soc.* **98**, 81–98 (2023).
3. S. Steinert-Threlkeld, *Philos. Sci.* **87**, 897–909 (2020).
4. J. R. Martin, thesis, Harvard University (2022).
5. T. Korbak, J. Zubek, J. Rączaszek-Leonardi, arXiv:2010.15058 [cs.NE] (2020).
6. S. Engesser, A. R. Ridley, S. W. Townsend, *Proc. Natl. Acad. Sci. U.S.A.* **113**, 5976–5981 (2016).
7. M. Leroux *et al*., *Nat. Commun.* **14**, 2225 (2023).
8. S. W. Townsend, S. Engesser, S. Stoll, K. Zuberbühler, B. Bickel, *PLOS Biol.* **16**, e2006425 (2018).
9. J. Kuhn, S. Keenan, K. Arnold, A. Lemasson, *Linguist. Inq.* **49**, 169–181 (2018).
10. P. Schlenker *et al*., *Theor. Linguist.* **42**, 1–90 (2016).
11. M. Leroux, *Rev. Primatol.* **14**, 16469 (2023).
12. J. P. Trujillo, J. Holler, *Sci. Rep.* **14**, 2286 (2024).
13. U. Sauerland, *Theor. Linguist.* **42**, 147–153 (2016).
14. P. Schlenker *et al*., *Biol. Rev. Camb. Philos. Soc.* **99**, 1278–1297 (2024).
15. P. Schlenker, C. Coye, M. Leroux, E. Chemla, *Biol. Rev. Camb. Philos. Soc.* **98**, 1142–1159 (2023).
16. G. J. L. Beckers, M. A. C. Huybregts, M. B. H. Everaert, J. J. Bolhuis, *Front. Psychol.* **15**, 1393895 (2024).
17. Z. S. Harris, *Word* **10**, 146–162 (1954).
18. G. Boleda, *Annu. Rev. Linguist.* **6**, 213–234 (2020).
19. F. Husson, S. Lê, J. Pagès, *Exploratory Multivariate Analysis by Example Using R*, Computer Science & Data Analysis Series (CRC Press, ed. 2, 2017).
20. T. Furuichi, in *Bonobo and Chimpanzee*, Primatology Monographs (Springer, 2019), pp. 1–36.
21. M. C. Corballis, *Cognition* **44**, 197–226 (1992).
22. T. N. Suzuki, D. Wheatcroft, M. Griesser, *Nat. Commun.* **7**, 10986 (2016).
23. K. Ouattara, A. Lemasson, K. Zuberbühler, *PLOS ONE* **4**, e7808 (2009).
24. P. Schlenker, E. Chemla, K. Zuberbühler, *Trends Cogn. Sci.* **20**, 894–904 (2016).
25. K. E. Langergraber *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **109**, 15716–15721 (2012).
26. G. Dezecache, K. Zuberbühler, M. Davila-Ross, C. D. Dahl, *R. Soc. Open Sci.* **4**, 160816 (2017).
27. L. S. Oña, W. Sandler, K. Liebal, *PeerJ* **7**, e7623 (2019).
28. C. Hobaiter, R. W. Byrne, K. Zuberbühler, *Behav. Ecol. Sociobiol.* **71**, 96 (2017).
29. C. Wilke *et al*., *Anim. Behav.* **123**, 305–316 (2017).
30. J. G. Mine *et al*., *Behav. Ecol. Sociobiol.* **78**, 108 (2024).
31. M. Fröhlich, C. P. van Schaik, *Anim. Cogn.* **21**, 619–629 (2018).
32. E. Genty, Z. Clay, C. Hobaiter, K. Zuberbühler, *PLOS ONE* **9**, e84738 (2014).
33. K. Liebal, K. E. Slocombe, B. M. Waller, *Ethol. Ecol. Evol.* **34**, 274–287 (2022).
34. K. E. Slocombe, B. M. Waller, K. Liebal, *Anim. Behav.* **81**, 919–924 (2011).
35. M. Berthet, M. Surbeck, S. W. Townsend, Dataset and R code - Extensive compositionality in the vocal system of bonobos, figshare (2025); https://doi.org/10.6084/m9.figshare.c.7648628.

ACKNOWLEDGMENTS

We thank the people of the villages of Bolamba, Yete, Yomboli, Yasalakose, and Bekungu, who granted us access to their forest; the Ministry of Research in the Democratic Republic of the Congo (DRC); the Institut Conglais pour la Conservation de la Nature; the Bonobo Conservation Initiative; and Vie Sauvage. We also thank the staff of the Kokolopori Bonobo Research Project, as well as L. Zanutto and M. Rohée, for their help and support during data collection. We thank G. Mesbahi and D. da Cruz for logistic support, as well as E. Ahmad, L. Bierhoff, C. H. Chen, L. Fornof, G. Gordon, and J. Vlaeyen for help and support during data collection. We also thank F. Wegdell for conducting an interobserver reliability test. We thank K. Slocombe, M. Leroux, and S. Watson for their comments on the draft and C. Zulberti, A. Bosshard, I. Schamberg, N. Lahiff, and J. Kuhn for insightful discussions. **Funding:** This work was funded by Swiss National Science Foundation grants PP00P3_198912 and 315130_192620 (S.W.T.), NCCR Evolving Language, Swiss National Science Foundation Agreement #51NF40_180888 (S.W.T.), and Startup from Harvard University (M.S.). **Author contributions:** Conceptualization: M.B., S.W.T.; Methodology: M.B., S.W.T., M.S.; Data collection: M.B.; Funding acquisition: S.W.T., M.S.; Supervision: S.W.T., M.S. (fieldwork); Writing – original draft: M.B., S.W.T.; Writing – review & editing: M.B., S.W.T., M.S. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** The dataset analyzed in the current study and the statistical script are available in the figshare repository (*35*).

SUPPLEMENTARY MATERIALS

science.org/doi/10.1126/science.adv1170
Materials and Methods
Figs. S1 to S4
Tables S1 to S9
References (*36*–*63*)

Submitted 6 December 2024; accepted 26 February 2025
10.1126/science.adv1170